自然言語処理 116－1
（1996. 11. 18）

複数辞書の統合的利用のための汎用日本語辞書の構築

羽田 ゆかり　　松本裕治
奈良先端科学技術大学院大学

本研究では、複数の辞書を相補的に用いて、計算機による自然言語処理のための汎用的な辞書を構築する方法について考察する。対象辞書として、EDR 電子化辞書、分類語彙表、JUMAN 辞書などの電子化辞書を用いる。この手法では、まず、様々な電子化辞書の品詞分類を比較・検討し、学校文法の品詞分類に自動的に分類し直すシステムを構築する。そして、複数の辞書を相補的に用いるような、自然言語処理のための汎用辞書を構築する。本稿では、この手法の基礎となる概念、原理、応用結果について述べる。

Integration of Multiple Lexicon and Construction of General Purpose Japanese Lexicon

Yukari Hada　　Yuji Matsumoto
Nara Institute of Science and Technology

This paper reports on the implementation of General Purpose Japanese Dictionary(GPJD) for the NLP systems from multiple lexicon. The EDR Japanese Word Dictionary, the Japanese Morphological Analyzer JUMAN Dictionary and the Bunruigoihyou Thesaurus are used as target lexica. In this method, several lexica are compared and contrasted in order to make transfer tables of part-of-speech categories. Next, the lexica are automatically classified in the traditional part-of-speech categories(Gakkou Bunpou). Then, the GPJD for the NLP system is constructed which integrates multiple lexicon.

1. はじめに

日本語などの自然言語を計算機で解析する自然言語処理では、その過程において単語辞書・シソーラス・文法規則などの語彙情報に関する様々な知識が参照され、解析結果の品質はこれらの語彙情報の質と量に影響される。この語彙情報は辞書という形で保存されているため、辞書は自然言語処理において重要な位置を占める。

自然言語処理で使用している辞書は、電子化辞書とも呼ばれ、単語の意味の定義だけでなく品詞情報・形態素情報・構文情報など単語に関するあらゆる情報を含むものである。従来から存在する国語辞典などの辞書は人間の膨大な知識や高度な推論・言語理解能力の下で使われることを前提に作成されている。しかし、自然言語処理システムでの利用を目的とする電子化辞書では、必要とされる情報を曖昧さのない明確な形式で記述しなくてはならない。

一般に良い電子化辞書の条件は、網羅性、汎用性、頑健性などが挙げられる。CD-ROM 版の国語辞典から EDR 電子化辞書のような自然言語処理システムのための辞書まで様々な辞書が存在するが、このような条件を全て満たすものは数少ない。現在特に問題となっているのが語数の不足と情報の欠落である。文法情報の充実した辞書、語彙の充実した辞書など各辞書には長所・短所があり、それらを複数使用すると足りない情報が補えるようになると考えられる。例えば、語彙の充実した辞書から文法を抽出・結合すると、語彙と文法共に充実した辞書ができる。

電子化辞書は複数種開発されているので、複数

の辞書を利用することが原理的には可能となっている。しかし、実際には形式の違う複数の辞書を組み合わせて使用するシステムはほとんどない。また複数の辞書を相補的に利用するには、それぞれの辞書の構造の相違点を解消する必要がある。

そこで本研究では、これらの辞書を一旦中間形式に変換することにより複数の辞書を自由に組み合わせて使用できるシステムを作成した。作成した辞書の有効性を検証するために、既存のシステムにこの辞書を適用し実験・評価を行った。

本稿では、まず2章で本研究が対象とする電子化辞書について、構造および特徴を把握する。3章では、汎用文法体系について説明し、2章での調査結果をふまえ、文法体系間の対応について述べる。4章では、3章では扱えなかった語彙の分類されていない電子化辞書をコーパスを用いて品詞分類する方法について述べる。これらに基づいて5章では、汎用辞書実現について述べる。そして、汎用辞書作成方法と辞書の品質に関する評価を行う。6章では、本手法の問題点についてふれ、今後どのように拡張していくかについてその展望を述べる。

2. 電子化辞書

2.1 EDR 電子化辞書

EDR 電子化辞書[EDR95]プロジェクトは1986年に開始され、約10年間に渡り試験研究を続け、日本語単語辞書、英語単語辞書、概念辞書、日英対訳辞書、英日対訳辞書、日本語共起辞書、英語共起辞書、専門用語辞書の8種類の辞書を完成させた。

単語辞書は日常一般に用いられる基本的な単語、一般化した専門用語や固有名詞、略語、そして熟語や成句などが収録されている。概念辞書には約38万にのぼる概念を体系上に分類した概念体系辞書と、格関係を中心に概念間に成り立つ関係を集めた概念記述辞書に分かれている。さらに、言葉の表層的な共起関係を記述した共起辞書や日本語と英語の単語間の対訳関係を記述した対訳辞書がある。これらの5種類の辞書は、互いに関連しあって、全体の辞書を構成している。

2.2 分類語彙表

分類語彙表[国立国語研究所 93,96]の仕様は、国立国語研究所報告 4『婦人雑誌の用語』(1953年)および同報告 13『総合雑誌の用語』後編(1958年)に収めた分類語彙表の分類の仕様に大体基づいている。1964年に出版された『分類語彙表』だけでは、収録語数が少なすぎるため、1981年から増補作業を始め、現在に至る。1996年3月現在で、収録語数は、87743 語である。分類語彙表は大きくわけて 4 つの分類(体の類、用の類、相の類、その他の類)から成り立っている。他の辞書と比べると意味情報中心なため、詳細な品詞分類はされていない。また、各分類が、抽象関係、人間活動主体、人間活動、生産物、自然などの項目に細分類されている。

2.3 日本語形態素解析システム JUMAN

システム部分と文法・辞書部分が独立した日本語形態素解析システム JUMAN[松本 94]の辞書は、品詞分類の定義、活用型の一覧、活用形の一覧、連接関係の定義、品詞辞書から構成されている。システム標準文法は、益岡・田窪文法[益岡 92]に基づいており、品詞体系を活用する語(動詞、形容詞、判定詞、助動詞)と活用しない語(名詞、副詞、助詞、連体詞、感動詞、指示詞)に分けている。JUMAN 辞書で使用されている益岡・田窪文法は、活用体系の呼び名や形容動詞を認めない立場にあることが特徴である。

3. 汎用辞書作成のための文法体系

3.1 汎用文法体系

品詞体系の作成や、それぞれの単語についての品詞の認定は非常に難しい問題である。そこで既存の文法をもとに分類することにした。本校では、IPA コーパス 95 年版[橋本 95]の文法[1]が、学校文

[1] IPA コーパス 96 年版[井佐原 96]では、品詞分類の方法が時枝文法に基づいており、学校文法には基づいていないため、本研究では採用しなかった。

JUMAN 辞書	IPA 95 年版
動詞 母音動詞	動詞 一段
動詞 子音動詞カ行一段	動詞 カ行イ音便

表 1 IPA95 年版文法から JUMAN 辞書へ変換する表 (一部)

法に基づいた品詞分類であるため、この文法を汎用文法体系として使用する。また、利用者の研究目的にあわせて自由に文法を変更して利用できるように、本システムでは表 1 のように IPA コーパス 95 年版の文法と利用したい文法体系の対応表を与えるだけで、利用者の望む文法体系に変換できる形となっている。

IPA コーパス 95 年版は、学校文法に基づいており、第 1 レベルとして、名詞・動詞・形容詞・形容動詞・副詞・連体詞・感動詞・助詞・助動詞・記号・その他を挙げている。これらの第 1 レベル 12 種の品詞に対して、必要に応じて第 5 レベルまでを表示した。

活用する語 (動詞、形容詞、形容動詞、助動詞) は、語幹と語尾をつなげて、それを一単位としてある。また、「勉強＋する」のような漢語サ変動詞については、語幹部分を「名詞 サ変接続」とし、「する」だけを動詞として扱っている。形容動詞に関しては、語幹部分が名詞、語尾部分が助動詞であるという可能性も高いため、語幹と語尾をつなげず分割した形となっている。

実際の自然言語処理システムでは、付属語に対する処理がすでに組み込まれている場合が多い。したがって、本研究で開発する汎用辞書では付属語を対象とせず、自立語だけから成る辞書を構成することにした。

3.2 文法体系間の対応

上記の IPA コーパス 95 年版の文法と EDR 電子化辞書、日本語形態素解析システム JUMAN 辞書の 3 つを比較・対比し、品詞名の変換テーブルを作成した。IPA コーパスそれぞれで使用されている文法体系を比較・対比させ、変換テーブルを作成した。このテーブルをもとに、各辞書の品詞体系を汎用日本語辞書の品詞分類に変換する。

分類語彙表は分類方法が異なるため、この手法では対応を示すことができない。そのため別の手法をとることにした。この手法については、次章のコーパスを利用した品詞の推定の節で詳しく述べる。

4. コーパスを利用した品詞の推定および語彙の抽出

本研究で扱った分類語彙表の文法体系は学校文法を基にした IPA コーパス 95 年版の品詞分類とは異なるため、汎用文法体系に基づいた品詞分類に分類しなおす必要がある。本章では、品詞分類されていない辞書や言語データの見出し語の品詞推定を新聞記事などのコーパスを用いて行う手法について述べる。尚、本稿では分類語彙表を中心にコーパスを利用した品詞の推定方法を述べるが、分類語彙表のように大きく概念別に分かれているものだけでなく、まったく分類されていないようなものに対しても、ある程度の品詞分類ができるような方法である。

分類語彙表の文法体系は、体、用、相、その他と分かれている。一般的に、本体・本質を「体」、作用・活動を「用」としている。ここでいう体の類とは、概念を表わす語で、文法上の特色は、主語となりうること、活用がないことである。そこで、分類語彙表の体の類を名詞であるとした。

次に用の類だが、動作・作用を表している。今回の分類では、分類語彙表の用の類を動詞であるとした。ここには、「副詞＋する」「形容詞＋がる」「動詞＋られる」と助動詞を伴う単語や動詞性接尾辞である「おいて」なども含まれている。

相の類の文法上の特色は、物事や動作を修飾するものである。分類語彙表の相の類を形容詞、形容動詞、副詞、連体詞のいずれかであるとした。

その他の類は、概念と概念との間の、また、叙述と叙述との間の関連づけや、感動や呼びかけや応答や、判断・期待・過程などの叙述態度や表現意図についての告知や待遇をあらわす語を含んでいるという定義から、接続詞、感動詞、そして、上記の三つの分類のどれにも属さないものを含ん

でいると考えた。

分類語彙表に記載されている語には、次のように分類番号が与えられている。

面会 1.3520　　求める 2.3520　　しっとり 3.3513　　とかく 4.3120

分類番号の整数一桁目は、「1. 体の類」「2. 用の類」「3. 相の類」「4. その他」の四つの品詞的特性[2] に基づいて分類されている。

分類語彙表からは、EDR 電子化辞書と日本語形態素解析システム JUMAN 辞書に含まれていない単語だけを抽出して分類した。実際にどのようにして品詞の分類を行ったかを各類に分けて説明する。

4.1 体の類

体の類は名詞であるとした。前述のように分類語彙表の体の類は「1.」で始まるため「1.」で始まるものだけを抽出して名詞とした。

4.2 用の類

前述のように、分類語彙表の用の類は「2.」で始まるため、「2.」で始まるものだけを抽出し動詞とした。動詞には、サ変接続や「副詞+する」なども含まれるため、見出し語を「〜する」とそれ以外に分けた。「〜する」以外のものについては、活用型を確定する必要があるため、コーパス中の活用変化を検出し、活用型を推定した。

サ変名詞も「副詞+する」も「する」で終わっているため、「する」で終わる単語だけを切り出し、「する」の前の単語がサ変名詞か副詞かをチェックする。これは、EDR と JUMAN の辞書上で、名詞であれば、サ名詞、副詞であれば 副詞とする。副詞でも「する」を取らない語もあるため、その他の情報として、「する」を取り得る語だと明記しておく。

さて、動詞には活用があるが、動詞の活用分類については、語尾が「る」のものについて、コーパスを利用して分類することにした。尚、語尾が「う、く、す、つ、ぬ、む、ぶ、ぐ」については、それぞれ五段活用「ワ行、カ行、サ行、タ行、ナ行、マ行、バ行、ガ行」と分かっているため、コーパスでチェックする必要はない。語尾が「る」の語については、以下の法則で分類した。

「〜る」の未然形

(A) 〜がア・ウ・オ段

→ くる　(1) こない → カ行変格活用

　　(2) くらない → ラ行五段活用

→ する　(1) しない → サ行変格活用

　　(2) すらない → ラ行五段活用

→ ずる　(1) じない → ザ行変格活用

　(2) ずらない → ラ行五段活用

→ 他はラ行五段活用

(B) 〜がイ・エ段

　(1) 〜らない → ラ行五段活用

　(2) 〜ない → 一段活用

語尾が「る」の語は、上の法則を利用して、未然+連用の形を毎日新聞 4 年分のコーパスに一度でも出現したらその活用であると決める。尚、毎日新聞 4 年分に全ての動詞が使われているとは思えないため、活用が分からない単語については、候補をあげるようにする。

4.3 相の類

分類語彙表の相の類では、形容詞、形容動詞、副詞、連体詞が主に扱われている。相の類をこれら 4 つの品詞に分類する方法だが、まず、形容詞

品詞	EDR	分類語彙表	JUMAN	汎用辞書 (正解率%)
名詞	189042	55443	149459	227992 (100)
動詞	44556	21669	14133	28107 (100)
形容詞	2161	9890	6471	2413 (100)
形容動詞	7937			8520 (99.99)
副詞	4634		2190	5763 (99.99)
連体詞	368		96	582 (97.81)
接続詞	378	1116	107	340 (100)
感動詞	574		91	741 (100)
その他	8444		439	
分類不能				5889
計	258094	88118	172986	279448 (98.28)

表 2 汎用辞書と電子化辞書間の品詞体系の対応と語彙数

[2] 厳密な品詞分類が行われていない場合もある。

と形容動詞は、活用するが、副詞、連体詞は活用しないことがわかっている。また、形容詞と形容動詞の語尾は、「い、な」であることから、活用する語と活用しない語に分けてから分類する。

活用する語は、「い」で終わっていれば形容詞の活用変化させ、「な」で終わっていれば形容動詞の活用変化をさせ、その単語が一度でもコーパスに出現すれば、それぞれ形容詞・形容動詞であると判定する。

活用しない語は単語の特徴が明らかなものを抽出し、その特徴を基に品詞推定を行う抽出に関しては、コーパスの観測に基づいて以下に示すような経験則を適用した。

* 「〜やか」「〜らか」の語を形容動詞とする
* カタカナ語を形容動詞とする
* こそあど言葉を連体詞とする
* 同じ音を繰り返す語を副詞とする
* 「〜の」「〜た」を連体詞とする
* 「〜に」「〜と」「〜り」を副詞とする

4.4 その他の類

その他の類では、接続詞・連体詞・副詞を扱っており、それ以外のものは扱っていない。語彙表には概念見出しがついているため、概念見出しによって接続詞と感動詞を抽出することが可能である。それ以外のものは副詞とした。

5. 実験と評価

本章では、(1)変換テーブルを用いて品詞体系を汎用文法体系に変換する、(2)コーパスを利用して品詞を推定する、(3)単語の特徴を用いて品詞を推定する、の三つの手法を実装し、本手法の有効性を検証する。

5.1 汎用辞書作成の実験

3 章で述べた手法に基づきそれぞれの辞書の記述内容を汎用辞書の体系に変換し、汎用辞書を構築する実験を行った。表 2 には、実際に抽出できた語数を示す。汎用辞書の欄は、EDR 電子化辞書・分類語彙表シソーラス・JUMAN 日本語形態素解析辞書から抽出した語を汎用辞書形式に変換し、それらをマージした結果である。尚、正解率については、5.3 節で述べる。

5.2 品詞推定制度の評価

4 章で述べた分類語彙表の見出し語の品詞推定の手法の検証結果を表 3, 4, 5 に示す。分類に関しては、コーパスを用いて品詞推定したものと単語の特徴を利用して品詞推定したものとに分けられる。

5.3 汎用辞書作成手法の評価

本節では、汎用辞書作成の手法について評価した。表 2 に示した正解率について述べる。EDR 電子化辞書・分類語彙表・JUMAN 辞書の名詞は、貧し変換テーブルが一意に決まるため、正解率を 100%とした。また、EDR 電子化辞書と JUMAN 辞書の動詞・形容詞・形容動詞・副詞・連体詞・接続詞・感動詞は、品詞変換テーブルが一意に決まるため、正解率を 100%とした。分類語彙表の動詞(一段活用・五段ラ行)・形容詞・形容動詞でコーパス中の活用変化検出により可能な活用型が一意に決まるものは、正解率 100%とした。分類語彙表の相の類の活用しない語と活用する語の一部は、経

品詞	出現数	未出現数
動詞	2435	629
形容詞	328	321
形容動詞	224	73

表 3 コーパス (新聞) を利用した品詞推定の結果

品詞	抽出源	出現数	未確定数
名詞	用	8354	497
名詞	相	233	0
副詞	用	21	名詞未確定数 497 に含まれる

表 4 コーパス (辞書) を利用した品詞推定の結果

品詞	特徴	正解数	出現数	正解率(%)
形容動詞	〜やか、〜らか	69	70	98.57
副詞	〜に	122	122	100
副詞	〜と	51	52	98.08
副詞	畳語	454	454	100
連体詞	〜た	181	191	94.76
連体詞	〜の	93	98	94.90

表 5 単語の特徴を利用した品詞推定の結果

月	JUMAN 辞書に含まれる自立語(%)	JUMAN 辞書以外の辞書にだけ含まれる自立語(%)	付属語(%)	未定義語(%)	出現語数
1	60.55	1.85	35.38	1.77	8304
2	62.47	1.75	34.36	1.42	7730
3	62.12	1.70	33.84	1.64	8129
4	61.61	1.87	34.65	1.86	8325
5	63.63	1.75	32.60	2.02	8393
6	62.65	1.79	33.76	1.90	7784
7	63.08	1.76	33.34	1.82	8243
8	62.72	1.66	33.77	1.85	8489
9	62.90	1.77	33.77	1.56	8086
10	63.01	1.75	33.67	1.57	8284
11	62.81	1.75	33.67	1.77	8067
12	61.78	1.89	34.36	1.70	8079

表 6 朝日新聞 1985 年度社説の形態素解析結果

験則に基づき品詞を推定した。これらは、人手 により正解率をチェックした。この正解率は 98.28% であった。分類語彙表のその他の類に出現する接続詞。感動詞・副詞も正解率を人手でチェックした。この正解率は 100%であった。分類不能であったものが 1.72%あった。したがって、この作成された汎用辞書の全体の正解率は 98.28%となった。

5.4 汎用辞書の品質評価

本節では、本研究で作成した汎用辞書の品質を評価した。具体的な手法として、汎用辞書を JUMAN 辞書の形式に変換し、1985 年度朝日新聞社説 1 年分を日本語形態素解析システム JUMAN で解析するのに使用した。そして、形態素解析の結果から基の JUMAN 辞書に含まれる自立語、JUMAN 以外の辞書にだけに含まれる自立語、付属語、未定義語の割合を調べた。その結果を表 6 に示す。この結果から、未定義語の出現率はわずかながら減ったことが分かる。この実験により本研究で用いた変換テーブルによる品詞体系変換はおおむね満足の行く結果であった。

6. おわりに

本稿では、複数の辞書を統合的に利用することが可能となるように汎用辞書を作成した。汎用辞書作成過程において得られた成果は次の通りである。まず、各辞書で使用している文法体系間の変換テーブルを作成して、異なる品詞体系に変換することが可能であることが示せた。そして、品詞分類のされていないデータに関しては、コーパスを用いることにより品詞推定ができその手法が有効であることを示せた。また、コーパスを用いることができなかったデータについては、語の特徴を用いて品詞推定を行いその手法が有効であることが示せた。さらに、作成した汎用辞書を用いて形態素解析を行った結果、未定義語の出現頻度が減った。

参考文献

[EDR95] 日本電子化辞書研究所、EDR 電子化辞書使用説明書(1995).

[国立国語研究所 93] 国立国語研究所 (編) : 分類語彙表、秀英出版(1964, 1993).

[国立国語研究所 96] 国立国語研究所(編) : 分類語彙表、秀英出版(1996).

[井佐原 96] 井佐原均、萩野紫穂、桑畑和佳子、徳永健伸、橋本三奈子、元吉文男 : RWC テキストデータベース報告書、報告書、技術研究組合新情報開発機構(1996).

[橋本 95] 橋本三奈子, 萩野紫穂、徳永健伸、元吉文男、井佐原均 : IPA コーパス概要、IPAL シンポジウム'95 論文集、pp. 31-44(1995).

[益岡 92] 益岡隆志、田窪行則 : 基礎日本語文法-改訂版-、くろしお出版(1992).

[松本 94] 松本裕治、黒橋禎夫、宇津呂武仁、妙木裕、長尾眞 : 日本語形態素解析システム JUMAN 使用説明書 version2.0、京都大学工学部長尾研究室、奈良先端科学技術大学院大学松本研究室(1994).